Engineering Data Analysis with Matlab

Anke Scherb

Engineering Risk Analysis Group
Technische Universität München

Phone: 089 289 23013
E-mail: anke.scherb@tum.de

Engineering Data Analysis with Matlab

Today’s lecture

- Multiple random variables
- Functions of random variables
- Monte Carlo simulation (MSC)

Multiple random variables

Models of uncertain quantities that are observed simultaneously, e.g.:

- Wave height and wave period
- Mechanical properties of the same material
- Load acting on structure and capacity of the structure

The individual random variables are gathered in a vector

$$\mathbf{X} = [X_1, X_2, \ldots, X_n]^{\mathrm{T}}$$

Multiple random variables

Joint CDF

e.g. for two random variables X, Y

$$F_{XY}(x, y) = \Pr[(X \leq x) \cap (Y \leq y)]$$

Note:

- The joint CDF is a non-decreasing function in each argument
- The joint CDF has limits $F_{XY}(-\infty, y) = 0$, $F_{XY}(x, -\infty) = 0$, $F_{XY}(\infty, \infty) = 1$

Multiple random variables

Discrete random variables – Joint PMF

e.g. for two discrete random variables X, Y

$$p_{XY}(x, y) = \Pr[(X = x) \cap (Y = y)]$$

Normalization rule: $$\sum_{\text{all } x_i} \sum_{\text{all } y_i} p_{XY}(x_i, y_i) = 1$$

Continuous random variables – Joint PDF

e.g. for two continuous random variables X, Y

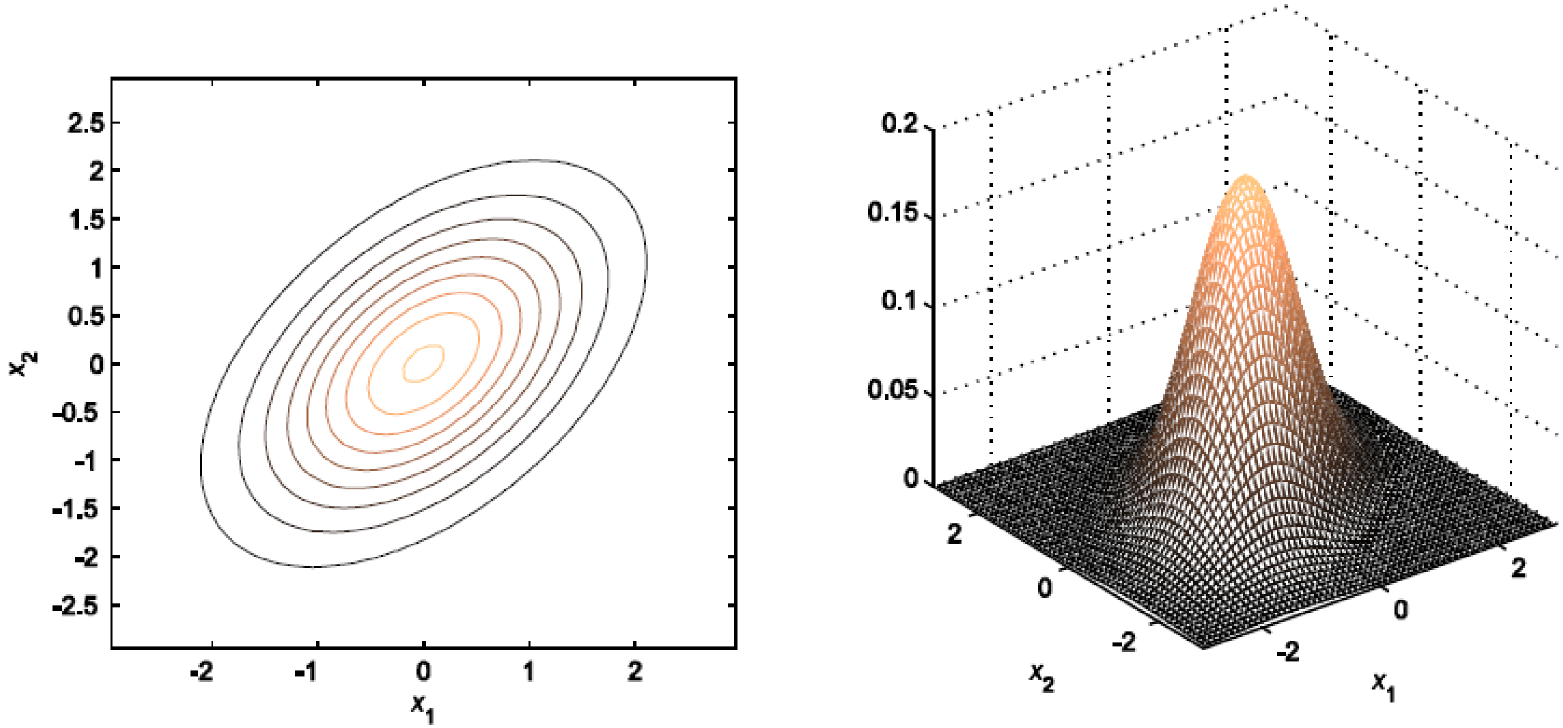
$$f_{XY}(x, y)dxdy = \Pr[(x < X \leq x + dx) \cap (y < Y \leq y + dy)]$$

Normalization rule: $$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx\, dy = 1$$

Multiple random variables

Example for two continuous random variables

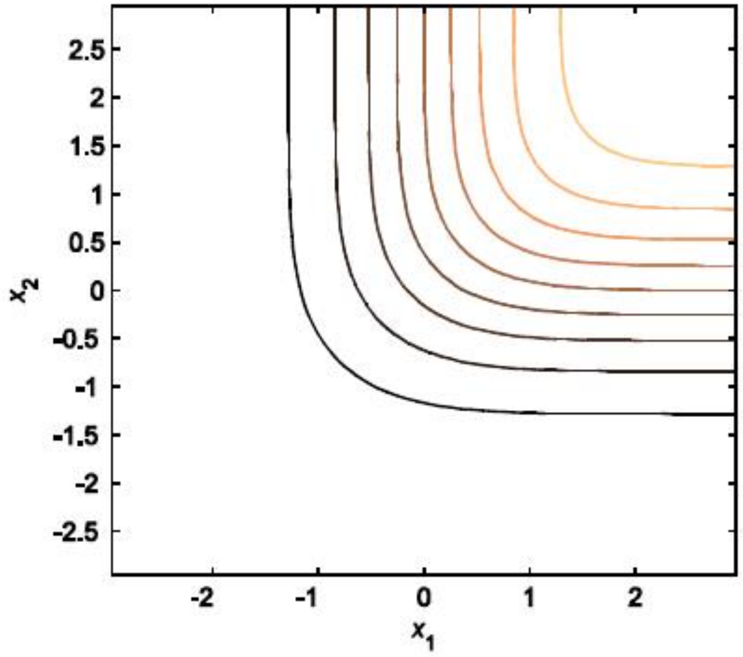
Joint PDF

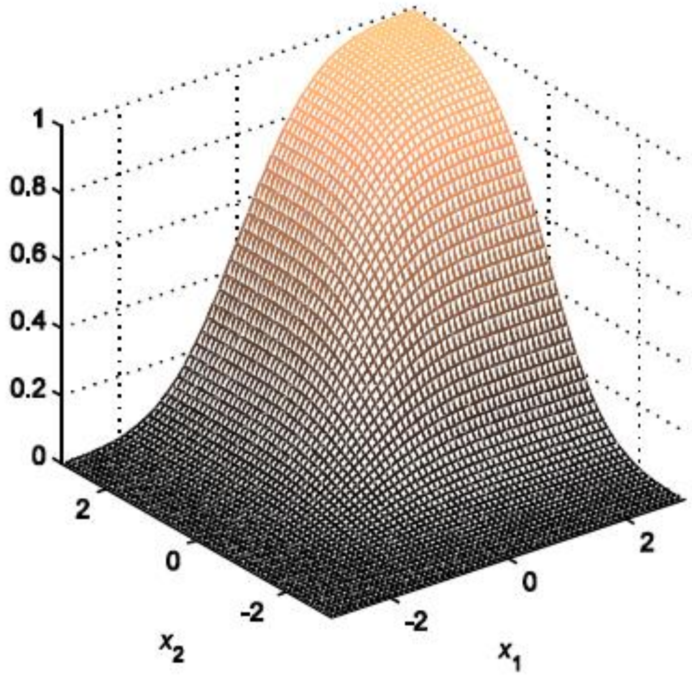


Multiple random variables

Example for two continuous random variables

Joint CDF





Statistical Independence

Proof of statistical independence of two events:

$$\Pr[(X = x) \cap (Y = y)] = \Pr(X = x)\, \Pr(Y = y)$$

For two S.I. discrete random variables X, Y

$$p_{XY}(x, y) = p_X(x)\, p_Y(y)$$

For two S.I. continuous random variables X, Y

$$f_{XY}(x, y) = f_X(x)\, f_Y(y)$$

Covariance

Mean vector

$$\mathbf{M}_{\mathrm{x}} = \mathrm{E}[\mathbf{X}] = [\mu_1, \mu_2, \dots, \mu_n]^{\mathrm{T}}$$

Covariance

$$\mathrm{Cov}[X, Y] = \mathrm{E}\left[(X - \mu_X)(Y - \mu_Y)\right]$$

- Two discrete random variables

$$\mathrm{Cov}[X, Y] = \sum_{\text{all } x_i} \sum_{\text{all } y_i} (x_i - \mu_X)(y_i - \mu_Y)\, p_{XY}(x_i, y_i)$$

- Two continuous random variables

$$\mathrm{Cov}[X, Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y)\, f_{XY}(x, y)\, dx\, dy$$

Covariance matrix and Correlation

Covariance matrix

$$\sum\nolimits_{\mathrm{XX}} = \begin{bmatrix} \mathrm{Var}[\mathrm{X}_1] & \mathrm{Cov}[\mathrm{X}_1,\mathrm{X}_2] & \mathrm{Cov}[\mathrm{X}_1,\mathrm{X}_n] \\ \mathrm{Cov}[\mathrm{X}_1,\mathrm{X}_2] & \cdots & \mathrm{Cov}[\mathrm{X}_2,\mathrm{X}_n] \\ \vdots & \ddots & \vdots \\ \mathrm{symmeric} & \cdots & \mathrm{Var}[\mathrm{X}_n] \end{bmatrix}$$

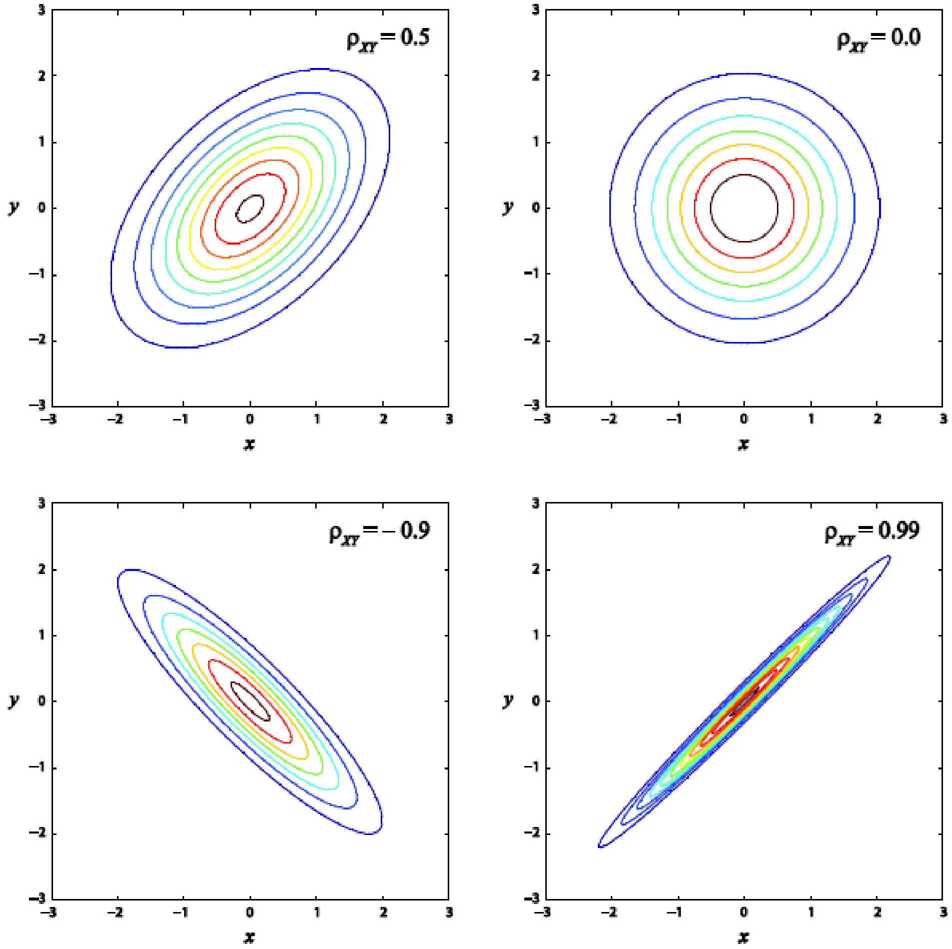
Correlation coefficient

$$\rho_{XY} = \frac{\mathrm{Cov}[X,Y]}{\sigma_X \sigma_Y}$$

Note: Covariance and correlation coefficient measure the linear dependence between two random variables

Description of random vectors

Bivariate normal distribution with varying correlation coefficients



Matlab – 3D Plotting

Plotting commands

`meshgrid`	Creates a rectangular grid in 2D or 3D
`mesh`	Plots colored parametric mesh
`surf`	Plots colored parametric surface
`contour`	Plots a contour plot
`meshc`	Combines **`mesh`** and **`contour`**
`surfc`	Combines **`surf`** and **`contour`**

Example: plotting3D.m

Functions of random variables

Engineers use models to describe physical, chemical, economical processes

- Models for flood prediction
- Models for evaluating deformations of a structural system
- Models to estimate the advancement of a construction process
- …

Models describe input-output relations

Input → $\boxed{\textbf{Model}}$ → **Output**

$\mathbf{X}$ $\qquad$ $g(\mathbf{X})$ $\qquad$ $\mathbf{Y}$

If $\mathbf{X}$ are random variables then $\mathbf{Y}$ will be random variables

Expected value of a function

- Discrete random variables $\mathbf{X} = [X_1, X_2, \ldots, X_n]^{\mathrm{T}}$

$$\mathrm{E[Y]=E[g(X)]=} \sum_{\text{all } x_1} \sum_{\text{all } x_2} \ldots \sum_{\text{all } x_n} g(x)p(x)$$

- Continuous random variables $\mathbf{X} = [X_1, X_2, \ldots, X_n]^{\mathrm{T}}$

$$\mathrm{E[Y]=E[g(X)]=} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} g(x)f(x)\, dx_1 dx_2 \ldots dx_n$$

Note:

- If $g(X) = X$, the expected value is the mean of X
- If $g(X) = (X - \mu_X)^2$, $\mathrm{E}[g(X)]$ is the variance of X

Expected value of a function

- Linearity of the expectation

$$\mathrm{E}[c] = c$$

$$\mathrm{E}[ag(\mathbf{X}) + c] = a\,\mathrm{E}[g(\mathbf{X})] + c$$

$$\mathrm{E}[g_1(\mathbf{X}) + g_2(\mathbf{X})] = \mathrm{E}[g_1(\mathbf{X})] + \mathrm{E}[g_2(\mathbf{X})]$$

wherein a and c are deterministic constants.

Expected value of a function

Example: Expected burnt area in a wildfire

- The *diameter* of burnt areas after wildfires in a particular region is exponentially distributed with mean value 100 m.
- What is the expected value of the *burnt area*, if the area is idealized by a circle?

Expected value of a function

Example: Expected burnt area in a wildfire

- The diameter D of burnt areas after wildfires is exponentially distributed with parameter $\lambda = 1/\mu_D = 10^{-2}\ \mathrm{m}^{-1}$
- The burnt area is idealized by a circle

$$A = g(D) = \frac{\pi}{4} D^2$$

Exponential PDF $f(d) = \lambda \exp(-\lambda d)$

Expected burnt area: $\mathrm{E}[g(D)] = \mathrm{E}[\frac{\pi}{4} D^2] = \int_0^\infty \frac{\pi}{4} d^2 \lambda \exp(-\lambda d)\, dd$

Example: wildfire.m

Distribution of a function of random variables

Derivation of the distribution of **Y** = g(**X**) when the distribution of **X** is known:

- For each value of X, there is a corresponding value of Y
- To find the distribution of Y, the inverse function of g(X) is needed:

$$X = g^{-1}(Y) = h(Y)$$

- The inverse function is often not available in analytical form

=> numerical algorithms often represent efficient and practical solutions

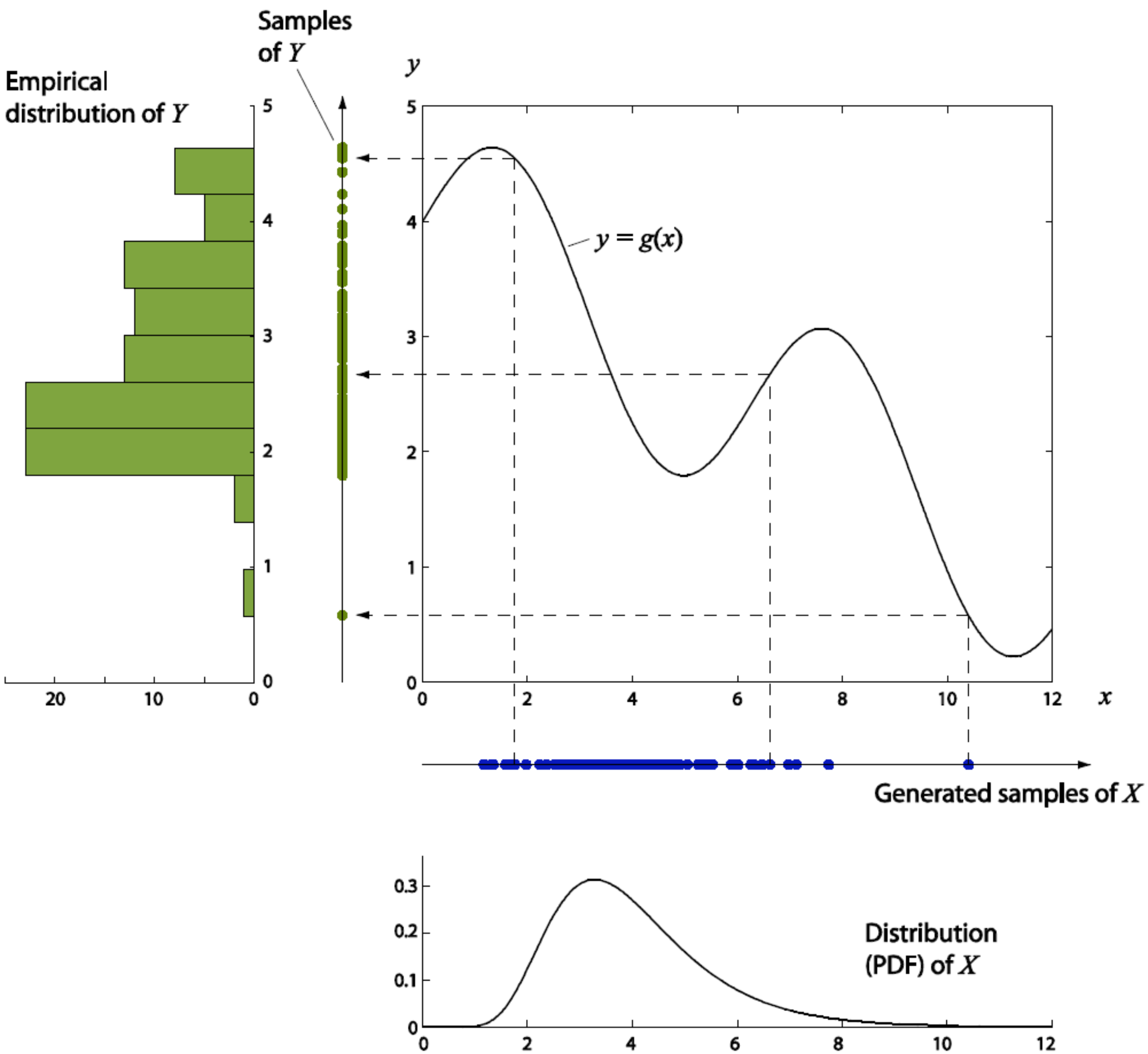
=> popular strategy: **Monte Carlo simulation**

Monte Carlo simulation

Steps

- Generation of random samples $\mathbf{x}_i$, $i = 1,\ldots,n_s$ of the input variable(s) $\mathbf{X}$
- Evaluation of the function at the sample values: $\mathbf{y}_i = g(\mathbf{x}_i)$
- Analysis of the generated samples $\mathbf{y}_i$ of $\mathbf{Y}$

Monte Carlo simulation



Monte Carlo simulation

- Generation of (pseudo-) random samples

  Pseudo-random number generators produce samples from the uniform distribution in $[0, 1]$

```
rand(m,n)
```

Pseudo-random: a computer cannot generate true randomness. Generated numbers are deterministic sequences that must be initiated by the so-called *seed* value

Monte Carlo simulation

- Generation of samples from a random variable with CDF $F_X(x)$
  - Generate a sample u_i uniformly distributed in $[0, 1]$
  - Require that the samples u_i and x_i have the same CDF value

$$F_U(u_i) = F_X(x_i)$$

$$u_i = F_X(x_i) \Leftrightarrow x_i = F_X^{-1}(u_i)$$

Assuming a strictly increasing CDF $F_X(x)$

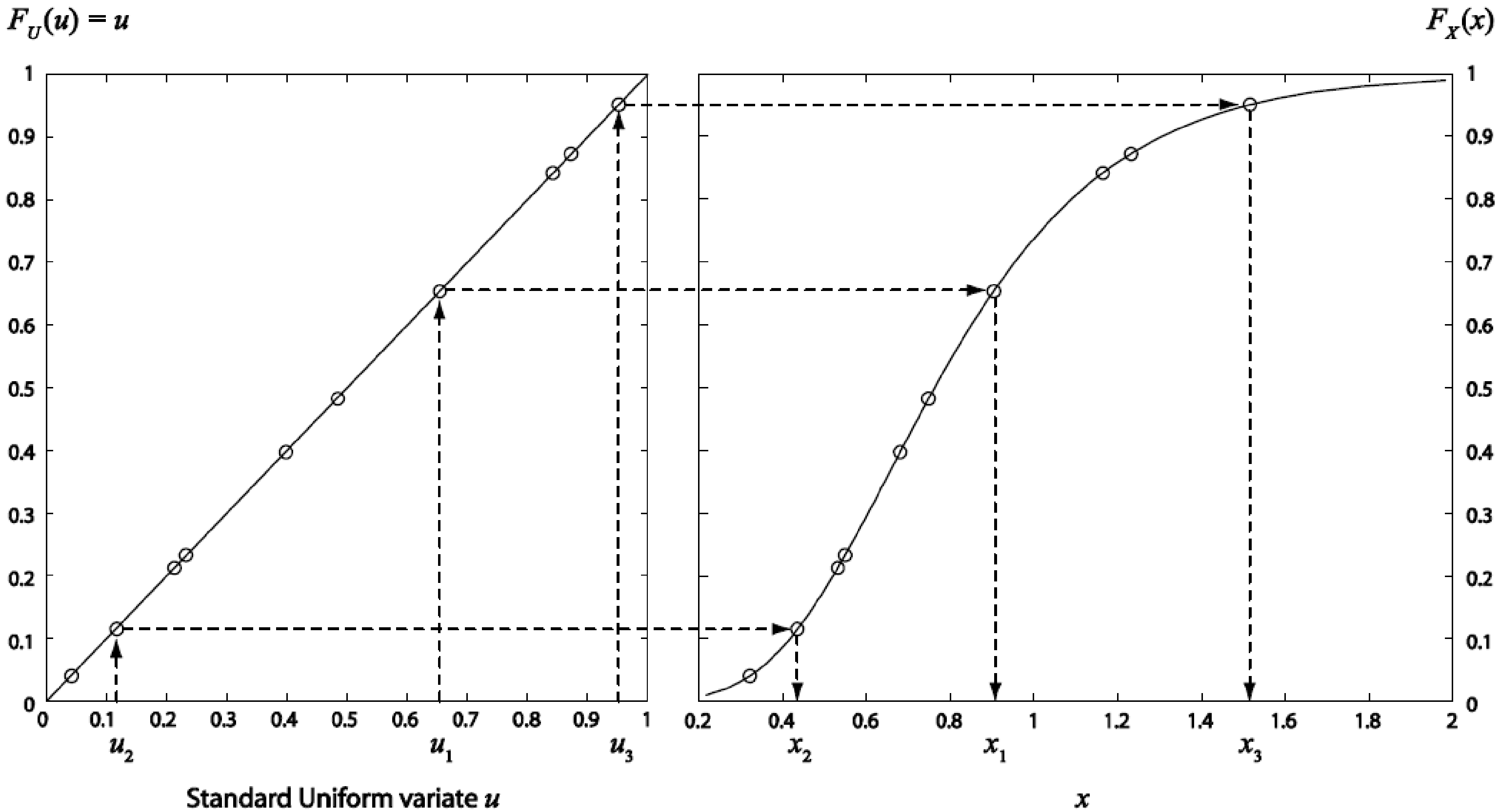
```
namernd(par1,par2,m,n,…)
```

or using a distribution object **`pd`**

```
random(pd,m,n,…)
```

Monte Carlo simulation

- Generation of samples from a random variable with CDF $F_X(x)$



Monte Carlo simulation

- Evaluation of the function

Input $\longrightarrow$ **Model** $\longrightarrow$ **Output**

$\mathbf{x}_i$ $\qquad$ $g(\mathbf{x})$ $\qquad$ $\mathbf{y}_i$

Repeating evaluations of a numerical model is commonly the most computationally demanding part of MCS

Monte Carlo simulation

- Analysis of the samples $\mathbf{y}_i$
  - Compute statistics (sample mean, sample standard deviation,…)
  - Plot graphical summaries (histograms, cumulative frequency diagrams,…)
  - Expected value of a function:

$$
\begin{aligned}
\mathrm{E}[g(\mathbf{X})] &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\mathbf{x}) f(\mathbf{x})\, dx_1\, dx_2 \cdots dx_n \\
&\approx \frac{1}{n_s} \sum_{i=1}^{n_s} g(\mathbf{x}_i) \\
&= \frac{1}{n_s} \sum_{i=1}^{n_s} \mathbf{y}_i
\end{aligned}
$$

Note: In MCS the multi-fold integral is replaced with single summation!

Monte Carlo simulation

- Analysis of the samples $\mathbf{y}_i$

Example for a typical problem: computation of a probability $\Pr(\mathbf{Y} \in \Omega)$, where Ω is a domain in the outcome space of $\mathbf{Y}$, e.g. $\Omega = \{Y \geq y_{\mathrm{cr}}\}$

$$\begin{aligned}
\Pr(\mathbf{Y} \in \Omega) &= \int_{\Omega} f_{\mathbf{Y}}(\mathbf{y})\, dy_1\, dy_2 \cdots dy_n \\
&= \int_{\Omega} I[\mathbf{y}_i \in \Omega]\, f_{\mathbf{Y}}(\mathbf{y})\, dy_1\, dy_2 \cdots dy_n \\
&\approx \frac{1}{n_s} \sum_{i=1}^{n_s} I[\mathbf{y}_i \in \Omega]
\end{aligned}$$

where

$$I[\mathbf{y}_i \in \Omega] = \begin{cases} 1, & \mathbf{y}_i \in \Omega \\ 0, & \text{else} \end{cases}$$

(Indicator function)

Monte Carlo simulation

Example: Reliability of a structure

- Limit state function of two random variables R and S:

$$g(R,S) = R - S$$

capacity demand

- Failure of the system $F = \{g(R, S) \leq 0\}$
- Probability of failure

$$\Pr[g(R,S) \leq 0] \approx \frac{1}{n_s} \sum_{i=1}^{n_s} I[g(r_i, s_i) \leq 0]$$

Monte Carlo simulation

Example: Reliability of a structure

- Assume a system where demand, S and capacity, R are normally distributed with:

$$\mu_R = 100, \sigma_R = 10, \mu_S = 50, \sigma_S = 12.5.$$

- Probability of failure: count samples that fall into the failure domain and devide by the number of samples

Example: reliability.m

Accuracy of Monte Carlo simulation

Example: Use MCS to estimate the expected value $\mathrm{E}[Y]$

- The MCS estimate for $\mathrm{E}[Y]$ is equal to the "sample mean":

$$\mathrm{E}[g(X)] = \mathrm{E}[Y] = \bar{Y} = \frac{1}{n_s}\sum_{i=1}^{n_s} y_i$$

- The sample mean is also a random variable:

$$\mathrm{E}[\bar{Y}] = \frac{1}{n_s}\sum_{i=1}^{n_s}\mathrm{E}[Y_i] = \frac{1}{n_s}\sum_{i=1}^{n_s}\mu_Y = \mu_Y$$ [Unbiased estimator]

$$\mathrm{Var}[\bar{Y}] = \frac{1}{n_s^{2}}\sum_{i=1}^{n_s}\mathrm{Var}[Y_i] = \frac{\sigma_Y^{\ 2}}{n_s}$$

Accuracy of Monte Carlo simulation

Example: Estimate of the expected value $\mathrm{E}[Y]$

- Standard deviation of the estimate

$$\sigma_{\mu_Y, MCS} = \frac{\sigma_Y}{\sqrt{n_s}}$$

- Coefficient of variation (standard deviation divided by mean) of the estimate

$$\delta_{\mu_Y, MCS} = \frac{\sigma_Y}{\mu_Y \sqrt{n_s}} = \frac{\delta_Y}{\sqrt{n_s}}$$

Accuracy of Monte Carlo simulation

Example: Estimate of the probability of failure $\Pr(\mathbf{Y} \in \Omega)$

- The MCS estimate is equal to the sample mean of $Z = I[\mathbf{Y}_i \in \Omega]$

- Assuming that p is the true value of the probability $\Pr(\mathbf{Y} \in \Omega)$, the mean and variance of Z are

$$\mu_Z = 1 \cdot p + 0 \cdot (1 - p) = p$$

Binomial distribution

$$\sigma_Z^2 = p - p^2$$

Accuracy of Monte Carlo simulation

Example: Estimate of the probability of failure $\Pr(\mathbf{Y} \in \Omega)$

- The standard deviation of the MCS estimate:

For small p_mcs

$$\sigma_{MCS} = \frac{\sigma_Z}{\sqrt{n_s}} = \sqrt{\frac{p - p^2}{n_s}} \approx \sqrt{\frac{p_{MCS} - p_{MCS}^2}{n_s}} \approx \sqrt{\frac{p_{MCS}}{n_s}}$$

- Coefficient of variation (standard deviation divided by mean) of the estimate

$$\delta_{MCS} \approx \frac{1}{\sqrt{p_{MCS} n_s}}$$

- Required number of samples for a target δ_{MCS}

$$n_s \geq \frac{1}{\delta_{MCS}^2 p_{MCS}}$$

Estimated probability of failure

Example: reliability.m